2007 年度後期　統計データ解析Ｂ　第１３回

データの背景にある構造を探るー因子分析

※このプリントは，広島大学外に公開している「学外用」プリントのため，他の書籍から引用した例を割愛しています．

第１０回の講義で，「見かけ上の相関」をとりあげました．このときの例では，「体格」と「成績」の相関が実は見かけ上のものであり，その相関は，「学年」というもうひとつの量を考えると，「『学年』と『体格』との相関」「『学年』と『成績』との相関」の２つの相関で説明できることを示しました．

因子分析は，この例の「学年」のように「複数の量の間の相関が，それらに共通な量との相関によって説明できる」というモデルを考え，この「共通な量」ー「因子」と言いますーを見つける方法です．この方法は，心理学の分野での，「表に現れる人の行動」に「背後にある心理的要因」があるのではないか，という考えから発達したものです．

因子分析の計算は複雑なため，１回の講義で全部を説明することはできません．今回の講義では，将来因子分析を使ってデータ解析をする機会が来たときのために，その基本的な考え方を理解してください．

因子を用いたモデル

例えば，数学，英語，理科，国語という４つの科目で行う試験を考えてみましょう．これらの科目の成績の間には，正の相関があるという仮説が考えられます．つまり，ある科目の成績がよい人は，他の科目の成績もよいと想像できます．

因子分析では，このように４科目間の相関があると考えられるとき，「実は，１つの『学力』という量が背後にあって，『学力』とそれぞれの科目の成績とに正の相関があるので，このような４科目間の相関が現れる」と考えます．極端な場合，もしも一人の受験生の点数が４科目とも同じであれば，各受験生の成績は４科目で表す必要はなく，１つの「学力」というデータになります．そんなに極端な場合でなくても，４科目間の相関が強ければ，各受験生の成績は「学力」というひとつの量で，かなりの程度表現できるはずです．

この場合の「学力」のように，測定はしていないけれども，相関のある複数の量を一度に説明できると考えられる量を，因子（共通因子）とよびます．

また，数学と理科の成績の間には強い正の相関があり，英語と国語の成績の間にも強い相関があるという仮説も考えられます．このように２科目間の相関が２組あるならば，「各科目の成績は，２つの因子で表現できる」と考えます．この場合，元の４科目のデータにおける各受験生の成績は，２つの因子（例えば「数理能力」と「言語能力」）でほぼ表現できると考えられます．このように，たくさんの項目（すなわち変量）からなるデータを少数の因子を使って表現し，測定されたデータの背後にある「そのデータが現れる仕組みのモデル」を見つけ出す方法を因子分析といいます．例えば，４科目の成績が，１つの「学力因子」であらわせるというモデルを考えたとしましょう．すると，ある受験生（例えば受験番号 i の「i 君」）の成績は

（i 君の数学の成績）＝（数学への影響の度合）×（i 君の「学力因子」の点数）＋（誤差）
（i 君の英語の成績）＝（英語への影響の度合）×（i 君の「学力因子」の点数）＋（誤差）
（i 君の理科の成績）＝（理科への影響の度合）×（i 君の「学力因子」の点数）＋（誤差）

（i 君の国語の成績）=（国語への影響の度合）×（i 君の「学力因子」の点数）+（誤差）

とあらわせることになります．

ここで，(「学力因子」の点数）はもちろん各受験生で異なりますが，これは「i 君の学力」を表すものですから，i 君という 1 人の人についてはどの科目についても同じになっています．一方，((科目）への影響の度合）は，(「学力因子」の点数）が実際の各科目の成績にどのくらい影響するかを表しているもので，科目間では異なりますが，ひとつの科目についてはどの受験生にとっても同じになっています．

ここでいう（(科目）への影響の度合）をその科目の**因子負荷量**，(i 君の「学力因子」の点数）を i 君の**因子得点**といいます．各科目の因子負荷量を求めることで，「学力因子」がどの程度実際の各科目の成績に影響しているかを調べることができます．また，各人の因子得点を調べることによって，各人の潜在的な「学力」の評価ができます．実際に因子分析を用いる場面では，個人の評価よりも，測定されたデータについて研究者が考えたモデルの妥当性を検証することが多いため，因子負荷量を調べることが重視される場合が多いです．

複数の因子がある場合

では，共通因子が「学力」ただ 1 つなのではなく，「数理能力」と「言語能力」の 2 つあるモデルを想定した場合を考えてみましょう．この場合，i 君の成績と因子との関係は次のようになります．

（i 君の数学の成績）=（数理因子の数学への影響の度合）×（i 君の「数理因子」の点数）
+（言語因子の数学への影響の度合）×（i 君の「言語因子」の点数）+（誤差）

（i 君の英語の成績）=（数理因子の英語への影響の度合）×（i 君の「数理因子」の点数）
+（言語因子の英語への影響の度合）×（i 君の「言語因子」の点数）+（誤差）

（i 君の理科の成績）=（数理因子の理科への影響の度合）×（i 君の「数理因子」の点数）
+（言語因子の理科への影響の度合）×（i 君の「言語因子」の点数）+（誤差）

（i 君の国語の成績）=（数理因子の国語への影響の度合）×（i 君の「数理因子」の点数）
+（言語因子の国語への影響の度合）×（i 君の「言語因子」の点数）+（誤差）

だんだん複雑になるので，記号で書いてゆきましょう．i 君についての，各科目の成績と，各科目における誤差（**独自因子**といいます）を，それぞれひとつにまとめてベクトル $\boldsymbol{z}_i, \boldsymbol{e}_i$ で表すことにしましょう．つまり，

$$\boldsymbol{z}_i = \begin{pmatrix} z_i(\text{数学}) \\ z_i(\text{英語}) \\ z_i(\text{理科}) \\ z_i(\text{国語}) \end{pmatrix}, \ \boldsymbol{e}_i = \begin{pmatrix} e_i(\text{数学}) \\ e_i(\text{英語}) \\ e_i(\text{理科}) \\ e_i(\text{国語}) \end{pmatrix} \tag{1}$$

とするわけです．また，i 君の因子得点も，(i 君の「数理因子」の点数）と（i 君の「言語因子」の点数）をひとまとめにして，次のようにベクトル $\boldsymbol{f}_i$ で表すことにしましょう．

$$\boldsymbol{f}_i = \begin{pmatrix} f_i[\text{数理}] \\ f_i[\text{言語}] \end{pmatrix} \tag{2}$$

さらに，因子負荷量を，例えば（言語因子の数学への影響の度合）を a[言語](数学) と書くようにすると，

上の成績・因子負荷量・因子得点・誤差の関係は，次のような行列とベクトルの計算になります．

$$\begin{pmatrix} z_i(\text{数学}) \\ z_i(\text{英語}) \\ z_i(\text{理科}) \\ z_i(\text{国語}) \end{pmatrix} = \begin{pmatrix} a[\text{数理}](\text{数学}) & a[\text{言語}](\text{数学}) \\ a[\text{数理}](\text{英語}) & a[\text{言語}](\text{英語}) \\ a[\text{数理}](\text{理科}) & a[\text{言語}](\text{理科}) \\ a[\text{数理}](\text{国語}) & a[\text{言語}](\text{国語}) \end{pmatrix} \begin{pmatrix} f_i[\text{数理}] \\ f_i[\text{言語}] \end{pmatrix} + \begin{pmatrix} e_i(\text{数学}) \\ e_i(\text{英語}) \\ e_i(\text{理科}) \\ e_i(\text{国語}) \end{pmatrix} \tag{3}$$

(3) 式に現れる行列（因子負荷行列とよびます）を A で表すと，(3) 式の関係は，

$$\boldsymbol{z}_i = A\boldsymbol{f}_i + \boldsymbol{e}_i \tag{4}$$

という式で表せます．

データを調べることによってわかるのは，各受験生の成績 $\boldsymbol{z}_i$ だけです．さらに，科目の数ははじめからわかっており，共通因子の数はモデルを考えた時点で決めました．因子分析の計算をひとことでいうと，これだけのことしかわかっていないときに，(4) 式を満たす因子負荷量 A と因子得点 f_i を求めること，となります．

共通性

因子負荷量や因子得点を，たったこれだけのことしかわからない状況で求めるには，いくつかの仮定をする必要があります．

- 各受験生の成績 $\boldsymbol{z}_i$ は，どの科目についても，平均 0，分散 1 になるように標準化されているものとします．
- 因子得点の単位・尺度は自由に決められるので，因子得点 $\boldsymbol{f}_i$ も，どの因子についても，各々平均 0，分散 1 に標準化されているものとします．
- 独自因子得点 $\boldsymbol{e}_i$ は，どの科目についても，いずれも平均 0 とします．
- 各独自因子得点の分散を

$$d^2 = \begin{pmatrix} d^2(\text{数学}) \\ d^2(\text{英語}) \\ d^2(\text{理科}) \\ d^2(\text{国語}) \end{pmatrix} \tag{5}$$

  であらわすことにします．
- 異なる科目の独自因子どうしの間，および独自因子と共通因子との間は相関がないとします．
- さらに，ここでは，共通因子どうしも（つまり「数理能力」と「言語能力」の間で）相関がないとします（このような因子を直交因子といいます）．

さて，例えば数学の成績の，受験生全体での分散 $V(z(\text{数学}))$ を求めると，(3) 式から

$$z_i(\text{数学}) = a[\text{数理}](\text{数学}) \times f_i[\text{数理}] + a[\text{言語}](\text{数学}) \times f_i[\text{言語}] + e_i(\text{数学}) \tag{6}$$

ですから，

$$V(z(\text{数学})) = \{a[\text{数理}](\text{数学})\}^2 \times V(f[\text{数理}]) + \{a[\text{言語}](\text{数学})\}^2 \times V(f[\text{言語}]) + V(e(\text{数学})) \tag{7}$$

となります[1]．ここで，上の仮定から，どの科目の成績もその分散は 1 で，また因子得点の分散も，どの因子についても 1 です．また，独自因子得点の分散は (5) 式で表されています．これらを用いると，

$$1 = \{a[\text{数理}](\text{数学})\}^2 + \{a[\text{言語}](\text{数学})\}^2 + d^2(\text{数学}) \tag{8}$$

ですから，

$$1 - d^2(\text{数学}) = \{a[\text{数理}](\text{数学})\}^2 + \{a[\text{言語}](\text{数学})\}^2 \tag{9}$$

という関係があることがわかります．この両辺の値のことを，（「数学」という科目における）共通性といいます．数学の共通性は，数学の成績の分散のうち，ここで用いた２つの共通因子で何パーセントが表現できたかを表しています．共通性が大きいほど，その科目の成績は今考えている共通因子を用いたモデルでうまく表せていることを意味しています．

因子負荷行列の推定

さて，因子負荷行列 A を推定するには，もう少し工夫が必要です．(8) 式を，各科目について並べてみましょう．

$$\begin{aligned}
1 &= \{a[\text{数理}](\text{数学})\}^2 + \{a[\text{言語}](\text{数学})\}^2 + d^2(\text{数学}) \\
1 &= \{a[\text{数理}](\text{英語})\}^2 + \{a[\text{言語}](\text{英語})\}^2 + d^2(\text{英語}) \\
1 &= \{a[\text{数理}](\text{理科})\}^2 + \{a[\text{言語}](\text{理科})\}^2 + d^2(\text{理科}) \\
1 &= \{a[\text{数理}](\text{国語})\}^2 + \{a[\text{言語}](\text{国語})\}^2 + d^2(\text{国語})
\end{aligned} \tag{10}$$

また，異なる科目の成績どうしの共分散を考えてみましょう．仮定から，異なる共通因子得点の間の共分散が 0 で，異なる科目どうしの独自因子得点の間の共分散も 0 なので，例えば数学の成績と英語の成績の共分散 $Cov(z(\text{数学}), z(\text{英語}))$ を求めると，これらの項が打ち消され，

$$Cov(z(\text{数学}), z(\text{英語})) = a[\text{数理}](\text{数学})a[\text{数理}](\text{英語}) + a[\text{言語}](\text{数学})a[\text{言語}](\text{英語}) \tag{11}$$

となります．どの科目の組み合わせについてもこのようになります．

(10) 式，(11) 式を，すべての科目，すべての変量について組み合わせると，各科目の成績の分散共分散行列（ここでは，成績が標準化されているので相関行列）を R とするとき

$$R = AA' + D \ \text{ すなわち } \ AA' = R - D \tag{12}$$

$$D = \begin{pmatrix} d^2(\text{数学}) & & & 0 \\ & d^2(\text{英語}) & & \\ & & d^2(\text{理科}) & \\ 0 & & & d^2(\text{国語}) \end{pmatrix} \tag{13}$$

という関係があることが導かれます（A' は行列 A の転置行列（行と列を入れ替えた行列）をさします）．

因子負荷行列 A を求めるには，(12) 式を解かなければなりません．もし $R - D$ がわかっていれば，第１２回の「主成分分析」の講義で触れた「対角化」の手法を使って解くことができます．しかし，実際には D は未知ですから，反復法によって近似解を求める必要があります．その詳細については，ここでは省略します．

---

[1]受験生全体のデータから，ひとつの「分散」が求められますから，添字の i はなくなっています．

単純構造と因子軸の回転

ここまでの説明では，はじめから「数理因子」「言語因子」という名前の共通因子があるように話を進めてきました．しかし，話の出発点は「４科目の成績が２つの共通因子で表される」というモデルを考えたことだけで，２つの因子の名前は便宜上つけただけです．上の計算で，確かに「数理因子」「言語因子」のような「わかりやすい因子」が得られていれば，このモデルは各科目の成績をうまく表すよいモデルだと言えます．しかし，そのような「わかりやすい因子」が求められているのでしょうか？ 残念ながらそれは違います．なぜならば，上の計算には次のような性質があるからです．

行列 T を，$TT' = I$（単位行列）がなりたつ任意の行列[2]とします．ある因子負荷行列 A について，(4) 式の関係が成り立っているとしましょう．(4) 式は

$$\begin{aligned} \boldsymbol{z}_i &= A\boldsymbol{f}_i + \boldsymbol{e}_i \\ &= A(TT')\boldsymbol{f}_i + \boldsymbol{e}_i \\ &= (AT)(T'\boldsymbol{f}_i) + \boldsymbol{e}_i \end{aligned} \tag{14}$$

と変形できますから，A が (4) 式のモデルを満たす因子負荷行列ならば，AT も (4) 式のモデルを満たす因子負荷行列ということになります．つまり，因子負荷行列はひととおりには定まらないわけで，ここまでの計算で求めた因子負荷量が，本当に「数理因子」「言語因子」の因子負荷量なのかどうかはわからないのです．

そこで，上の計算で求められた因子負荷量 A から，「わかりやすい」共通因子の因子負荷量になっている AT をさがします．そのような AT をみつけた結果，それが「数理因子」「言語因子」になっていれば，最初の仮説は正しかったことがわかるわけです．T が直交行列のとき，AT は行列 A を構成する各行ベクトルを回転する変換になります．そこで，AT から「わかりやすい」共通因子をさがす操作を因子軸の回転といいます．因子軸の回転にはさまざまな手法があります[3]．

「わかりやすい」共通因子とは何かについては，いろいろな考えがありますが，そのひとつに「各共通因子が，少数の科目に対しては相関が高く，他の科目との相関は低い」というものがあります．つまり，どの共通因子が，どの科目とどの科目に影響しているかがはっきりしている，ということです．この状態を**単純構造**といいます．

単純構造と因子軸の回転の意味は，図１を見るとよくわかります[4]．この図の左は，各科目に対して２つの因子に対する因子負荷量を図で表現したものです．この状態では，因子１が４つの科目全部と相関が高く，因子の意味がよくわかりません．これに対して，図１の右のように因子を変換すると，因子１は「理科」「数学」の各変量と相関が高く，因子２は「国語」「英語」と相関が高いことがわかります．もしこのような共通因子がみつかれば，最初に考えたモデルで「『数理能力』と『言語能力』という因子がある」と言うことができます．

実際のデータ解析では問題はもっと複雑ですから，「数理能力」と「言語能力」といった意味のはっきりした因子が，はじめから想定できることはあまりありません．しかし，因子軸の適切な回転を行って，各科目に対応する因子負荷量を観察することで，各因子がどういう意味をもつのかを考えることができます．

ただ，忘れてはならないのは，このような各因子の解釈ができるのは，あくまで最初に考えた「いく

[2] このような行列を直交行列といいます．

[3] 私の講義「応用統計学」(2004 年度前期）第１３回で，すこし触れています．さらにくわしくは，次ページの参考文献を参照してください．

[4] 出典：長谷川「ホントにわかる多変量解析」pp. 176–177 より翻案．

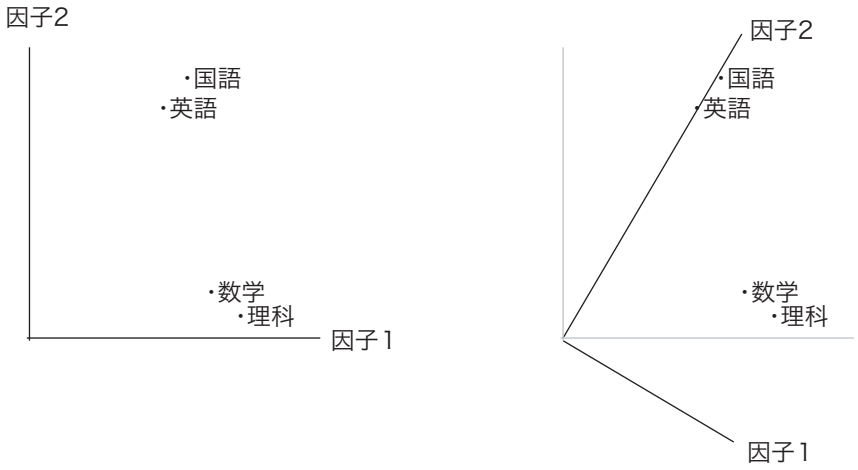

図 1: 単純構造と因子軸の回転

つもの科目の点数が，少数の因子で説明できる」というモデルを前提にしたうえでのことだ，ということです．因子分析の危険な点は，因子の解釈が「こじつけ」になりやすいことです．その解釈が本当に意味のあることなのかどうか，統計学的に，および固有科学的に，よく検討する必要があります．

因子分析についての参考文献

(上のものほど易しいものになっています)

大村平，多変量解析のはなし，日科技連　ISBN4-8171-2211-0
長谷川勝也，ホントにわかる多変量解析，共立出版　ISBN4-320-01591-6
田中豊・脇本和昌，多変量統計解析法，現代数学社　ISBN4-7687-0154-X
柳井晴夫・高根芳雄，現代人の統計２・新版多変量解析法，朝倉書店　ISBN4-254-12508-9

また，一部で因子分析に触れている次の本もおすすめできます．

東京大学教養学部統計学教室編，人文・社会科学の統計学，東京大学出版会　ISBN4-13-042066-6
松原望，計量社会科学，東京大学出版会　ISBN4-13-042069-0

(講義で用いたプリント（受講生にのみ配布）では，この後，田中・脇本「多変量統計解析法」p.p.189 より引用した例を掲載して，因子分析の実例を説明しています．)